CHAPTER XI

Testing: Intelligence, Aptitude, Personality, and Achievement[1]

G. M. RUCH and P. T. ORATA

INSPECTION OF SEVERAL HUNDRED TITLES in the field of testing, chiefly between January 1938 and July 1939, indicated that intelligence testing continued to hold first place in number of articles published within the scope of this summary, despite enormous gains in activity in the measurement of aptitude and personality. The bibliography of this chapter, therefore, represents a high degree of both quantitative and qualitative selection in order to place emphasis on new developments in measurement. The omission of much otherwise significant material rests solely on the basis that no outstanding uniqueness of method is involved.

Intelligence

The summaries to 1938 in the *Review of Educational Research* for June 1938 by P. Cattell (501) and Keys (561) comprising, respectively, 68 and 187 titles, need comparatively little supplementation so far as methodology and results are concerned. Developments since that date center chiefly in three areas: (a) further discussion of the 1937 L-M Scales of the Stanford-Binet, (b) renewed controversy over the constancy of the IQ arising chiefly from recent work at the state university of Iowa, and (c) the potentialities of factor analysis in psychological and educational measurement.

The New L-M Stanford-Binet Scales

The two books by Terman and Merrill (619, 620) presenting the L and M scales were the signal for renewed consideration of the Binet method, both for school use and for clinical purposes. Kent (559) presented suggestions for another revision, arguing that the age-scale method is wasteful and not well adapted to clinical practice, particularly when there is "nondiscriminative material at the upper and lower ends of the subject's natural range." R. B. Cattell (502) criticized the intuitive method in test construction in applied psychology, particularly the Binet tests. Vernon's reply (630) to Cattell and other critics of the Binet discussed the merits and limitations of both psychometric and clinical approaches, holding that the Binet embodies both methods. Both Cattell and Vernon listed numerous references. Bernreuter and Carr (487) and Merrill (572) discussed the significance of the IQ's yielded by the new scales; the latter pointed out that the addition of new tests at both the lowest and highest

[1] Bibliography for this chapter begins on page 610.

levels of the scale results in a different interpretation of the quotients on old and new scales. For example, the lowest 2 percent of the 1916 sample of 905 cases tested 73 or below, but the lowest 2 percent of the 2,904 cases used in the new scale tested 70 or below. A new table for the interpretation of L-M quotients is presented, with a disclaimer that classification as defective is possible by tests alone. Burt (496) compared the 1937 Binet with the English version of the 1916 scale. He found the new scale more effective in diagnosing the dull and defective, more reliable for showing the relative influences of a general versus other factors, but that the order of difficulty of subtests did not agree with that for English children.

The Constancy of the Intelligence Quotient

The nature-nurture controversy has received new attention as the result of a series of studies from the Iowa Child Welfare Research Station. Skeels (604), basing his conclusions on children tested 12 to 60 months after being placed in foster homes, reported in 1936 that "the mean level of intelligence of these children is higher than would be expected . . . from the educational, socio-economic, and occupational level represented by their true parents." He found zero correlation between true mother's IQ and child's IQ. Later studies by Skeels and others (602, 603, 605) reached the same conclusion. Wellman followed her 1932 report (639) of a steady rise in IQ year by year for children attending the Iowa Preschool Laboratories with four additional papers (634, 636, 637, 638) advancing similar conclusions. She said: "The extent of upward change that may take place is truly remarkable. We have examples of children entering preschool with average intelligence who, after especially favorable circumstances, have later tested at the 'genius' levels" (634). Stoddard (610, 611, 612) concluded that intelligence level is not fixed, but that a richer environment stimulates genuine mental growth. Skodak (606) and Crissey (509) also suggested that intelligence is much more responsive to environment than has previously been supposed.

This group of Iowa studies has been vigorously challenged by Simpson (600, 601) who claimed that the significance of these studies of the "wandering IQ" is completely obscured by ambiguities and inconsistencies in tabular data, failure to report individual scores year by year, and failure to allow for selective factors in school-leavers. He argued that the rises in IQ mean nothing more than "a survival of the fittest." Wellman (636) replied to Simpson's interpretations, and in another paper (635) she showed a similar, but less marked, increase in Merrill-Palmer IQ's under repeated testing.

Factor Analysis in Psychological Measurement

Cureton and Dunlap (511) summarized the work on the factor theory up to 1938 in the *Review* for June 1938, and Chapter XIII of the current

number, by Holzinger and Harman, brings the literature up to date. The publication of Thurstone's *Primary Mental Abilities* (622) provided the educator with a more concrete picture of the types of test materials that mathematical analysis suggests as useful in the measurement of primary mental traits. It is sufficient at this point to call attention to the fact that the methods of factor analysis and intuitive analysis of psychological abilities present fundamental differences. The time for the production of mental tests based upon factor analysis is at hand; just what similar analysis of educational abilities will yield by way of new achievement tests and curriculum reorganizations is at present challenging but purely speculative. Alexander (478) and Feder (527) applied factor analysis methods to educational tests and Guilford (538) to the production of four new forms of Army Alpha.

Books and Bibliographies

Outstanding aids to the test worker are: Buros' *1939 Mental Measurements Yearbook* (494) and Hildreth's revised *Bibliography of Mental Tests and Rating Scales* (551). The former provides (usually) two or more critical and independent reviews of tests published since the appearance of Buros' two earlier monographs. An innovation is the inclusion of reviews of books in the field of statistics and measurement. Educational, aptitude, and personality tests are also considered. Hildreth's bibliography is an extension and revision of her 1933 volume, with 4,279 titles classified under 17 headings. These two books provide virtually complete coverage of the field, and together constitute a working reference library of test materials. Revisions appeared of Freeman's *Mental Tests* (529) and Inglis's tables (554) of IQ values, which now provide the extensions demanded by the new Stanford-Binet scales.

Miscellaneous: New Tests, Revisions, Reliability, and Validity

Kuhlmann (564) presented a new intelligence scale resulting from his work with the Binet and Kuhlmann-Anderson tests. Norms were based upon 3,000 cases of ages from three months to adult. In addition to mental ages, scoring methods provided, above age nine, speed and accuracy ratings. The IQ was replaced by the PA (percent of average), as being less variable. Recent revisions are: Otis Quick Scoring (582); Detroit First-Grade Intelligence Test (523); and the Michigan Non-Verbal Series (536). Kerr (560) in England and Miller (573) in America continued to study the value of children's drawings for the measurement of intelligence. Higginson (550) published an objective test of imagination and Carl (499) devised a test for older children and adults in which a hole is filled with blocks of geometric forms; a reliability of .88 based on 1,508 adults and a correlation of about .77 with the Stanford-Binet were reported for the last mentioned test.

Strang (613) and Brill (492) discussed the validity of the Porteus maze test. Williams and Lines (641) evaluated the Ferguson form boards and derived new norms. Evaluations of other tests were reported as follows: Metropolitan Reading Readiness Test and Pintner-Cunningham Primary Mental Test, Grant (535); Kuhlmann-Binet, Arthur (481); Goodenough drawing test, McCarthy (569); CAVD tests, Pintner and Stanton (585); and the Spearman Visual Perception Test, Arsenian (480). Peatman (584) and Jackson (555) discussed the reliability and meaning of test scores.

Blatz and others (491) presented the growth of the Dionne quintuplets in a series of five monographs. Outstanding conclusions are: The girls show general retardation, especially in language; there are marked and more or less stable personality differences that appear to be environmental in character; and the quintuplets are more retarded in speech than a control twin group.

Watson (632) reviewed the intelligence movement, concluding that there is a great need for more studies of mental development. MacMurray (570) compared gifted and dull-normal children by the Pintner-Paterson and Binet scales. Wilson and Fleming (642) studied the intercorrelation of abilities in the first grade. Vernon (629) suggested that sophistication or test-wiseness may be an important element in test scores.

Aptitudes

Books and Reviews

O'Rourke (581) examined more than 500 studies in aptitude measurement and reviewed 130 vocational aptitude tests in the *Review* for June 1938. Outstanding books have appeared by Bingham (489) and Paterson, Schneidler, and Williamson (583). The former discussed the nature of aptitudes and the theory of their measurement. The latter is to be regarded chiefly as a handbook of directions, norms, and data on the validity and reliability of available aptitude tests, particularly those devised at Minnesota.

Aptitude for High School and College

Darley (512), Dickter (515), Langlie (565), and Selover and Porter (599) studied the use of psychological tests in predicting college success. Darley suggested that ability, attitudes, and college adjustment are probably unrelated. Dickter found that the mathematical parts of the CEEB examination predicted success in mathematics, but the verbal elements were of little value. Seagoe compared the predictive values of certain achievement tests with those of special aptitudes in algebra (596) and foreign languages (597). The New Stanford Achievement Test (arithmetic and reading) provided fully as good a basis for prediction as did the intelligence tests or special measures of aptitude in these subjects.

Aptitudes for the Professions and for Selling

Dwyer (521) gave the Strong interest tests to 418 entrants to medical school. Four factors (physicist, journalist, minister, and life insurance salesman) proved of most predictive value. Harris (543) administered five mechanical aptitude tests to 68 dental freshmen; these, combined with intelligence scores, gave a multiple correlation of .67 with dental school work. Stump (616), Stuit (615), and Sandiford and others (594) attempted to find predictive measures of teaching success. Aptitude tests, scholarship ratings, and success in practice teaching all proved to have low predictive value.

Lawe and Raphael (566), Dodge (516), and Bills (488) studied the values of tests for selecting salesmen. The former reported satisfactory results from tests employed at Harrods, Ltd. in London, and suggested the existence of upper and lower critical scores. Dodge listed nine items of the Bernreuter Personality Inventory that differentiate good and poor salesmen. Bills found that the life insurance selling and real-estate selling factors of the Strong interest test are related to success in selling insurance. Candee and Blum (497) devised a new scoring system for the Minnesota Clerical Test.

Mechanical Aptitudes

Drake (517, 518) and Drake and Oleen (519) evaluated various tests for selecting industrial employees and studied the psychological factors necessary to success on the job. Outstanding findings were: a new pin board with a reliability of .92 and a correlation of .59 with foremen's ratings; a new hand-foot coordination test; and 30 percent savings in operation through dual, or two-hand, operation. O'Connor (579, 580) presented further analysis of the Black Cube, Work Sample 167. Candee and Blum (498) gave the O'Connor finger dexterity and tweezer dexterity tests to mediocre and superior workers in a watch factory; the latter proved valid but the former was not except at a lower critical level. Age and experience did not affect the finger dexterity test. Wells (640) published his fourth paper on four O'Connor tests. Hearnshaw (548) described selection tests for inspectors in a paper mill. Burr and Metcalfe (495) revised the norms on the I. E. R. Assembly Test. Babcock and Emerson (482) analyzed the MacQuarrie mechanical ability test, reporting a correlation of .62 with the Binet vocabulary test—a correlation which increased with age as did the intercorrelations of the MacQuarrie subtests.

Personality, Interests, and Attitudes

Books and Reviews

Watson's summary (631) in the June 1938 *Review* of 329 titles published between January 1935 and December 1937 suggested the enormous

activity in the measurement of personality and character. Traxler (623) examined critically the leading tests in the field, listing 183 titles in his bibliography. Important books or monographs appeared by Thorpe (621), Murray (575), Garrett (533), Hartshorne and others (544), and Spencer (609). Thorpe's volume gave a comprehensive survey of the literature from all fields of psychology. Murray presented a detailed study of fifty men of college age over a two and one-half year period. Garrett used Thurstone's centroid method of factor analysis as a first step in defining and measuring personality traits. Hartshorne and others studied the intellectual, social, moral, and physical growth of 1,200 boys. Spencer's volume presented the personality conflicts of high-school students as revealed by a paper-and-pencil questionnaire.

Experimental Studies of Personality Tests

The Bernreuter Personality Inventory continued to receive critical attention. Farnsworth (526) retested 319 college students with the Bernreuter inventory at intervals of one, two, and three years. Responses proved relatively stable with time, as were intercorrelations. Jarvie and Johns (556) concluded that the Bernreuter inventory offers little aid in educational counseling. Nemzek (578) found the inventory of little value in predicting academic success as measured by teachers' marks. Hayes (547) decided that college women with several older siblings tended to be more neurotic and less self-sufficient and dominant, a finding previously reported by Stagner and Katsoff. Bennett (486) further simplified the Flanagan method of scoring the Bernreuter inventory.

Research on the Rorschach Ink Blot Test took mainly the direction of standardization of this clinical method. Troup (624) applied this test to twenty pairs of identical twins; no marked resemblances in temperament were found. Hertz (549) and Suares (617) attempted to objectify scoring and provide further norms. Fosberg (528) reported that the reactions were stable under retesting, even with changed directions. Rorschach data from many investigators were summarized by Davidson and Klopfer (513).

Interests and Attitudes

Strong (614) published a new edition of his Vocational Interest Blank for Men. Kopas (562) developed a "point-tally" method for scoring the Strong blank. Using the Strong scores as a criterion, Estes and Horn (525) constructed two scales that would differentiate between interests in mechanical and in electrical engineering. Carter and Jones (500) found the Strong scores to be closely related to high-school students' vocational choices. Garrison (534) and Cleeton (505) developed new interest inventories. Davies (514) cautioned test workers against giving interests an "all-determinative" role in vocational choices.

In an extensive study of 3,758 students in four state universities and fourteen church colleges, Nelson (577) gave the Lentz C-R Opinionaire to determine the prevalence of radicalism. The mean scores tended toward conservatism; few radicals were found; seniors were less conservative than freshmen; women were more conservative than men; and small differences were found from school to school. Corey and Beery (507) concluded that liking for school subjects is closely related to liking for the instructor.

Ratings

Few, if any, unique contributions in the use of ratings came to the attention of the reviewers within the period covered here. One extensive study was that of Eells (522) who studied the best liked and least liked aspects of 200 secondary schools, securing 24,000 returns. Scales were formulated that grouped these aspects under such headings as school staff, curriculum, pupil activity program, and guidance. (For other studies on ratings, see references 409, 493, 504, 574.)

Miscellaneous

Bell (485) published an adult form of his adjustment inventory that includes: (a) home adjustment, (b) health adjustment, (c) social adjustment, (d) emotional adjustment, and (e) occupational adjustment. Experimental studies of the following personality measures have been made, as follows: Willoughby (524), Baxter (484), Stanford M-F test (592), Woodworth-Cady and Baker "Telling What I Do" test (643), and the Loofbourow-Keys Personal Index (591).

Achievement

Books, Reviews, and Monographs

In the *Review* for December 1938, which summarized more than four hundred studies of educational tests and their uses, Scates (595) emphasized the changing conception of measurement, shown particularly in the work of the Eight Year Evaluation Study.

McCall's *Measurement* (568), a revision of his early text, was conspicuous for its shift toward the aims of the progressive education movement. Smith (607) devoted 182 pages to a critical examination of concepts of testing, a volume that has already proved very stimulating to those interested in the fundamentals of measurement. South (608) compiled a glossary of terms used in measurement and guidance. Ruch and Segel (593) prepared a handbook for counselors, on the use of the individual inventory in guidance. Segel (598) also summarized the cumulative record systems of 177 school systems.

Test Technics

Omitting for the present the Evaluation Study, several papers on test technics should be mentioned. May (571) wrote a very penetrating discussion of the logic of measurement. Kuder and Richardson (563) and Remmers and Whisler (590) considered critically the concept of the reliability coefficient, particularly its limitations. Kelley (558) showed that, under defined conditions, "upper and lower groups consisting of 27 per cent from the extremes of the criterion score distribution are optimal for the study (of the validity) of test items." Lev (567) used the method of analysis of variance to evaluate items and give them their proper weights. Guilford (539, 540) applied Fechner's law to the scaling of test items, holding that the easiness of an item is proportional to the logarithm of the magnitude of the stimulus. Dunlap (520) found that two-response tests requiring underlining were more open to scoring errors than certain other response forms.

Evaluation versus Measurement

Although logically a part of the discussion of achievement testing, this concluding section of the present summary is set apart for the sake of emphasis. It is a greatly condensed treatment of a review involving 129 titles. Space limitations have often necessitated the omission of authors' names in the citations.

The Philosophy and Function of Evaluation

The change in terminology from "testing" and "measurement" to "appraisal" and "evaluation" was regarded by Hosic (552) as a significant development in education. Tyler (627, 628) held that the emphasis is not on the relative merits of tests, but on the extent to which evaluation instruments promote as well as measure important outcomes of instruction. According to this point of view the functions of evaluation are no different from those of the school as a whole, namely, to help provide more intelligent guidance of teaching and learning, to develop more effective curriculums and educative experience, to secure more intelligent and effective cooperation with parents and community, and to provide an adequate and objective basis for measuring, recording, and reporting progress that facilitates the desired learning (576, 618).

The hypotheses of evaluation—According to Wrightstone (646) the new point of view in evaluation is based on a number of hypotheses radically different from those of Thorndike. Curriculum change and evaluation are coordinate aspects of the educative process; a program of evaluation should be comprehensive; and present instruments are inadequate for many of the major objectives of education. Hence there is need for a variety of new means and technics for gathering evidence. The measures should

correspond to the functional units of pupil behavior in actual curriculum situations; reliable and valid objective instruments of measurement are restricted to an appraisal of limited aspects of pupil behavior; and measures of functional behavior can best be developed by teachers working in cooperation with test technicians.

The Nature of Desired Achievement in the School Subjects

In view of these newer objectives of education the school subjects are expected to show correspondingly new kinds of results. Art should develop initiative, interest, judgment, and cooperation (537); the physical sciences, the ability to use experimental methods in gathering, organizing, and interpreting scientific data and in applying scientific facts and principles (483, 530, 531, 532, 541, 627); English literature, the reading of literature understandingly, a broader understanding of life, greater sensitivity to social problems, and increasing intelligence with regard to human motives and purposes (557); French and Latin, sufficient command of French and Latin vocabularies for simple reading and speaking (647, 650); home economics, better health, and a happier home life for all members of the family (503, 553, 586, 587); mathematics, thoroughness and precision in thought and action, disposition to question the validity of assumptions, expressed or implied, and sensitivity to the logic of arguments (545, 546); nursing, proper attitudes toward patients, other nurses, and physicians, and a wide range of interest not only in nursing but also in other and related fields, as well as the proper habits and skills in the performance of nursing activities (510, 626); social studies, sensitivity to and disposition and ability to deal with social problems in an intelligent manner, interest in international affairs and human welfare, and attitudes favorable to social improvement (479, 576, 589, 648, 651, 652); health and physical education, physical fitness, lively curiosity, self-confidence, and quickness and decisiveness of movement (508, 649).

Constructing Newer Achievement Examinations

The steps in constructing achievement examinations, according to the foregoing point of view, are likewise different from the well-established technics of objective test construction. They may be summarized as follows: (a) specifying the objectives of the school program as a whole; (b) restating, if necessary, each of these objectives in the light of the nature, characteristics, and requirements of the course, field, unit, or area in the school program that is to be evaluated; (c) defining the types of behavior that normally show whether or not and to what extent the objectives are being realized; (d) selecting test situations that will evoke the types of student behavior patterns consistent with the objectives; and (e) trying out these test situations with a view to improving their validity

and reliability and, at the same time, working toward making them more practicable (626, 627).

Evaluation instruments—A number of new instruments have been constructed both for appraising the school as a whole (542, 589, 644, 645, 646, 651) and specific segments or areas of it. Of the latter type are those that bear the following titles: A Scale of Beliefs, Interpretation of Data, Familiarity with Sources of Data, Application of Principles of Thinking (several subject fields), Interest Index, Problems Relating to Proof in Mathematics, Literary Information Test (both English and American literature), Questionnaire on Reading Interests and Reading Outcomes, Critical Mindedness in the Reading of Fiction, Judging Effectiveness of Written Composition, Questionnaire on Voluntary Reading, Descriptive Test Profile, Evaluation of Reading, and a Checklist of Magazines (588).

Developing a comprehensive program of evaluation—Tyler (576, 621) suggested ways and means of developing a program of evaluation that is both comprehensive and practicable "by making the appraisal an integral part of the learning process, by encouraging the pupil to make his own evaluation, by utilizing situations for evaluation which throw light upon the pupil's development at those points where the collection of direct evidence is highly impracticable."

Needed research in evaluation—In order to develop a comprehensive program of evaluation, research is needed "in discovering types of behavior which ought to be appraised, in devising means for appraising each important type of behavior, in refining appraisal instruments, in interpreting test results, and in follow-up studies regarding the permanence of learning" (625). Research in interest evaluation was also stressed by Weedon (633).

Critical Evaluation of Evaluation

Curiously enough, "evaluation" has been criticized on the same grounds as those on which it has criticized "measurement." The following represent some of the negative comments on the work of Tyler, Wrightstone, and others. The technics may be valid, but "not adequate"; "failure to supply clearcut data regarding the practices evaluated seriously weakens the scientific validity of the study"; "the study is subject to the usual lack of reliability, validity, and adequate sampling"; and "statistical significance alone does not prove educational significance" (494:271-72).